Consideration of Salmonids:

Consideration was given to sensitive species (e.g., salmonids) and rare and endangered species (e.g., sturgeon). Juvenile Chinook salmon are not as common in Suisun Marsh as they should be; they occur when temperatures are low and duck clubs are not as active in management, so the probability of juvenile salmon encountering harmful DO habitat conditions is limited at this time. As such, the rationale supports a suitable application of salmonid life history, monitoring and sensitivity data and science to ensure special protection during the period salmonids are likely to be present. However, restoration projects to improve salmonid habitat are ongoing and because the presence of juveniles can change in response to the impact of habitat restoration projects or, presumably, with climate change, this should be evaluated on an ongoing basis.

Assumption that Striped Bass is Sufficiently Sensitive to Represent Protection of the Fish and Benthic Macroinvertebrate Communities:

While striped bass is an introduced species, we note that they are an established species within the marsh fish community for over 100 years and are among the most sensitive species to dissolved oxygen levels in the marsh. Given that DO tolerance data for native species or other species that are possibly more sensitive were not available, the striped bass provided a suitable species for the analysis. In addition, Suisun Marsh is a "novel ecosystem," an ecosystem that is inhabited by an established community of native and nonnative species, and as such it is appropriate to include striped bass in the analysis, both in the calculation of proposed criteria as well as in the larval recruitment model. Most of the life stages for striped bass found in the marsh are juveniles and adults, life stages which have the ability to avoid poor DO conditions compared with egg and larval phases. The native species that has a prolonged larvae stage in the marsh is prickly sculpin (January-May), but data are not available to run a larval recruitment model with this species.

Use of Reference Study Data:

This study suitably employed available monitoring data on reference and impacted sloughs to assess the frequency of allowable excursions from derived criteria. The frequency of allowable exceedances should be based on the ability of aquatic ecosystems to recover from the exceedances, which will depend in part on the magnitudes and durations of the exceedances (U.S. EPA 2016). Exceedances are extreme values in the distribution of ambient concentrations and this distribution is the result of the usual variations exhibited by dissolved oxygen in the system. Because exceedances are the result of usual variation, most of the exceedances will be small and exceedances as large as a factor of two will be rare (U.S. EPA 2016).

This review and statistical assessment of the low and high temporal density time series of dissolved oxygen dynamics in reference and impacted sloughs has helped to ensure the criteria and their application are appropriate to the setting and informing the application of the criteria (e.g., monitoring, averaging period, exceedances). A comparison with reference conditions added insight into its application and the appropriateness of the VP approach. A CCC approach was used to capture the alternative exposure model used in the VP approach, and proved to be applicable and protective.

We note, however, that spatial data are limited and factors driving reference profiles can change rapidly. We cannot discount the possibility that characterization of reference conditions could be refined with the addition of data from other station locations. The suitability of these data for application to other habitats may not be appropriate.

Often times, the DO issues for water quality management are closely aligned with the warmest times of the year and the longest daylight hours. A conceptual model of the stressor-impact relationships affecting DO conditions facilitates targeting the critical periods and metrics for monitoring and management. At Suisun Marsh, there are multiple periods of concern that include the warm season, a season associated with growth and survival of rare, threatened and endangered species, and finally a period where the marsh conditions are affected by the impacts of culturally relevant activities (e.g. wetland management supporting suitable duck hunting conditions) that can introduce potentially harmful water quality conditions to the sloughs late in the year. The suite of considerations given to varied stressors in time throughout the year underlies the need for adaptive monitoring and management. Therefore, this criteria derivation process is not for a one-size-fits-all DO issue but considers targeting key seasons related to marsh management and the timing of sensitive species habitat use. The DO saturation levels and diel swings in the data might be indicative of nutrient enrichment, or effects of regulated marsh drainage and mixing of low DO waters, and diminish the utility and appropriateness of the data for setting system-wide, allowable criterion excursions. Nevertheless, until future work is done, the recommended excursions appear to be protective of the resource, only minimally deviating below the proposed criteria.

Existing Suisun Marsh non-probability spot survey data of fish abundance and water quality conditions (P. Moyle) represented an outstanding resource to support the approach by presenting an independent check on the derived criteria, as it presents a good idea of DO tolerance ranges. Additional monitoring, research or modeling would help confirm the representativeness of Mallard Slough as a "minimally impacted" system, as a natural condition, and relative to other habitats. It will also help to provide an independent check when future exceedances might be natural or related to climate change.

Use of DO Concentration versus Percent Saturation as Basis of Criteria:

Criteria should be as simple as is practical for supporting the management of the resource and communication among stakeholders. Oxygen concentration is perhaps our simplest common index of the DO conditions and is used effectively across the country to manage our aquatic resources. We acknowledge that saturation has had support in assessing water quality in San Francisco Bay (e.g. 3 months at or above 80% saturation) and Florida recently considered a DO saturation-based standard for their water quality assessments (P. Tango, Pers. Comm). Saturation can further provide useful diagnostic information on eutrophication and diel and seasonal DO patterns as to cause and whether the conditions are caused by natural or human drivers. It is worth moving the science and management world forward on how we might further use percent saturation in the criteria setting and assessment processes, because this measure is directly relevant to oxygen supply for fish and invertebrate respiration. In addition, climate change can bring direct influences on temperature and salinity, upon which concentration measures are dependent. Therefore, while not part of the criteria, saturation should be reviewed to assess natural excursions and for setting allowable exceedances from the criteria, specifically with regard to temperature and salinity.

Spatial Heterogeneity:

The need to consider spatial heterogeneity in Suisun Marsh is reasonable, given the well documented spatial heterogeneity of DO dynamics in marsh habitats (Boynton et al. 2014). In Suisun Marsh, small dead-end sloughs can naturally have low DO due to DOC loads from productive marsh and subtidal habitats. However low DO is rarely a problem in larger sloughs, due to mixing effects of wind, tides, and inputs of river water, especially during the critical summer season.

The study and its component analyses considered large and small sloughs, which was a reasonable level of classification. While segmentation of habitats supports site-specific DO criteria and standards assessments, consideration may be given to further spatial specificity in the study area given the addition of new data streams and sources. In addition to spatial variation, the vertical dimension warrants further evaluation to ensure that near-bottom conditions, which may affect benthos and demersal fish or life stages, are similar to monitoring data that have been taken at various depths.

Q2 Comment on the recommended averaging periods to assess attainment and DO evaluation, including:

*a. Averaging periods for CMC (**1-day** mean) and CCC (**30-day** mean), including step-wise approach to calculate daily and 30-day averages to compare against objectives;*

b. Should there be an allowable period of non-compliance of the CMC and CCC protective of aquatic life and estuarine communities, but not to be overprotective?;

c. Comment on whether, in back-end sloughs the CMC exceedances may occur multiple times in a month without adversely affecting the beneficial uses, as long as the CCC criteria are maintained;

d. Preferred monitoring window to detect the worst DO conditions (e.g. from mid-September through mid-November);

e. Comment on whether temperature and conductivity should be recorded together with DO to aid data interpretation and troubleshooting.

f. Preferred Monitoring Interval? 15-min?

Overview of Components Comprising the Water Quality Standard:

The averaging period cannot be divorced from several other critical aspects of the water quality standard: 1) minimum monitoring program station density, 2) minimum sampling frequency and type (discrete, continuous), 3) allowable frequency of non-compliance, and 4) the temporal averaging statistic. Per the proposed outline of criteria, “multiple samples” could use more explicit definition - How many stations will be monitored and need to pass the criterion/criteria to support a decision of attainment, OR, what is the assumption about how much water a single monitoring station represents, (i.e., what is the implicit interpolation of the results for how much water is represented by station results?).

Agreement will be needed on the criterion/criteria application periods. Presently those periods proposed appropriately address the critical periods of the warm season, the time period associated with actions linked to socially relevant activities, and rare and threatened species. Additional details of the sampling plan to support the water quality assessment should address the relevant depth or depths of sampling, location of samples or include the method for choosing sample sites (will samples be collected mid-channel, nearshore or both?). Boundaries to the criteria application should be detailed - do the criteria apply to some portion of the water with a minimum depth boundary of perhaps 1m or 0.5m? With respect to these considerations, here we review the recommendations and note areas where the recommendations could be strengthened to provide greater specificity for their linkage to the recommended monitoring program.

Averaging Period:

The averaging period for the CMC was shown to be effective for implementation of the criterion, both as a moving average and daily mean. Neither approach of block versus rolling is necessarily incorrect, what is important is that the method used to evaluate natural conditions and determine acceptable exceedances should be reflected in the assessment protocol. Such decisions will maintain the continuity for the basis for understanding the target status for the system based on a consistent understanding between characterizing the conditions and carrying out their assessment. Variances within a data stream are reflective of the sample intensity; the greater the sample density, the higher the probability of detecting outliers. If the assessment of reference conditions and likely exceedance rate of a criterion is, for example, 5% of daily means over a year based on 15 minute interval data, then monitoring protocols should attempt to be consistent with the behavior of the system as understood by also applying a 15 minute interval data collection for the assessment.

While the case was made that there were only small statistical differences between 7-day and 30-day static or block averages (i.e. stepwise) and rolling averages, data are somewhat limited and may not reflect the range of conditions that might be experienced in Suisan Marsh habitats. Use of the 30-day block averaging appears to provide a protective CCC under general conditions reflective of the data, but may not fully capture the combined impact of concentration and duration reflected in the larval recruitment model (Table 13 of the Tetra Tech report). Just as averaging period (i.e., 7 vs 30 day) in the block approach shows differences between the two in Table 17, depending on the starting and ending date for each "block", a 30-day block approach may miss the duration factor (i.e., allowable days) as the clock is "reset" after 30 days. However, the cumulative stress effect on aquatic life would, of course, continue into the next block but may be missed depending on when the consecutive blocks intersect the hypoxic period. Use of 7-day and 30-day moving averages were viewed as providing a more powerful statistical assessment of the conditions and, thus, better protection for aquatic life by assuring the durations assessed are indeed "continuous". As more locations are monitored and evaluated and more data become available, better understanding of condition variability by location, change over time (new sources, climate effects, management improvements), or other variations may be revealed that will guide and confirm appropriate monitoring and attainment protocols.

Minimum Requirements for Monitoring:

It will be helpful to state clearly the proposed assessment protocol which is integral to the monitoring program.

When developing a sampling plan, the basic question is whether or not the criterion is met. This is a binomial question where the foundations of the criteria derivation can be linked to the assessment method. Assume the reference data for a complete year show that the fish community is not stressed by low oxygen when no more than 5% of the daily means fall below the criterion. This would reasonably support a reference condition-based foundation for setting a binomial hypothesis to protect the related beneficial use. Without these supporting data, a 10% exceedance might be an allowable value acceptable to EPA if there were nothing else to go on. But in this case an assessment based on data is more supportable.

The criterion test should also be computed in the manner in which the criteria were derived. For example, if the daily mean criterion is computed from 15 minute data, sampling data should likewise be derived from 15 minute intervals. Every change in the sampling interval changes the variation that is captured. In this case, 5% of the conditions were known to be allowable based on the reference analysis, hence the 5% came about from variance in the data set collected at 15 minute intervals, which needs to be preserved in the sampling program.

There are at least two paths to consider given the stated criteria:

1. Use of a continuous monitoring sensor. Sampling continues for the entire year (or a specific target season). As described above, the assumption applied in this analysis is that this monitoring site is representative of the average conditions for the year in this habitat area, and a percent violation rate is computed. This assumption is reasonably supported since there is complete knowledge of system behavior based on a full year (or target season) of data. If the computed percent violation is below or equal to 5% then the waterway is presumed to meet the criterion. If the violation rate is more than 5% then that year fails its criterion. The continuous data represents comprehensive knowledge of the site conditions in time, and if the assumption of spatial representativeness for the area is supported, then also in space. (If another assumption is appropriate, e.g., a site is representative of "X" km^2 – for a local example, see Jassby et al. 1997 for an approach considering monitoring site representativeness in San Francisco Bay - then additional sampling sites could be required. For this example, it is assumed that one representative, continuous monitoring site to represent slough conditions is used.) This is not testing a hypothesis; a full accounting of the conditions is available so the output comparison is made directly against the criterion to support a statement about meeting or exceeding the criterion.

2. Limited resource and data assessments. Assume a daily mean criterion test is based on a continuous monitoring sensor deployed for 24 hours at a time, 12 days a year. In this case, some attention to randomization is needed. For example, a site or station is sampled beginning on a random starting date in January (e.g., Jan 17th) and then sampled the 17th of every month. Alternatively, a random date could be selected for sampling each month. Either way, the random element supports the statistical integrity underlying the tests. Further randomization could be used to select sites to address spatial variation concerns. So, data are collected on 12 random days, measuring every 15 minutes, 24 hours each sampling day. It is known from setting the criteria that the allowable exceedance rate is 5%. To test the sampling results against the allowable rate consider: With 12 sample dates, what is the probability that one daily mean in 12 will fail the criterion? What is the probability that your system experiences less than or equal to 5% violation rate of the criterion and that two daily means of your 12 fall below the criterion? Attachment 1 provides further details of applying this approach and understanding the likelihood of an event given an expected violation rate

ED_002551_00001479-00215

and translating the results to support statements of meeting or violating the criterion. A survey of State programs across the U.S. further shows the use a variety of monitoring approaches and decision criteria to support decisions on noncompliance with the criteria and nonattainment of a standard (Attachment 2).

Assessment Period to Determine Impairment:

Finally, regarding appropriate assessment periods that provide adequate grounds for a decision on impairment, understanding ecosystem recovery pace provides an important perspective (Attachment 2).

For example, a standards decision rule based on a 1 in 5 year allowable exceedance technically translates to an allowable exceedance of 20%, which does not constitute an impairment unless the violation rate is 40% (i.e., 2 or more years in a 5 year period). A "1 in 3" rule by comparison would mean the violation rate is 67% before an impairment is declared; however, the opportunity to take management action is nearer term than with a 5-year or longer period.

In related standards assessment work that considers decision rules for declaring impairments, under Chesapeake Bay criteria assessment protocols, a 1 in 6 years has recently been given consideration for future chlorophyll *a* assessments and provides for an allowable exceedance rate of 16.7%; 2 or more years out of attainment, or a minimum of 33.4% failure, equates to nonattainment (P. Tango, Pers. Comm.). By comparison, the State of Georgia uses a decision rule with a 5-year lake assessment period for chlorophyll *a* where 1 in 5 years moves the lake to Category 3 (i.e., Clean Water Act 303d listing classification that characterizes assessments as having insufficient available data and/or information to make a use support determination) and additional information is then used to evaluate attainment or impairment. For chlorophyll *a* assessment in Beaver Lake, Arkansas, Scott and Haggard (2015) suggested one alternative to assessing changes in average condition in 5-year windows may be to use a window as large as 10 years. The 10-year window was suggested to take into consideration decadal patterns associated with common climate cycles. Further, 1 year in 10 would be a 10% allowable exceedance. The adoption of a decision framework should further reflect an expectation that common recovery from impairments can occur and consider how attainment is tracked which can be in two forms – one for regulatory reporting requirements and one for a management tracking indicator.

Exceedance Frequencies:

There are two measures of exceedance that need to be considered. First is the criteria exceedance rate that equates to a violation. The second is associated with the decision on impairment – how many years can your system fail to meet criteria and over what time frame and still be considered in attainment?

This body of work conducted on Suisun Marsh data is probably supportive of "reasonable potential" protection for understanding an appropriate measure of allowable exceedances, but the exceedances are a measure of DO condition with respect to the criteria and may not always be indicative of supporting aquatic life beneficial uses. Using moving averages at two time scales for CCC will reduce potential for Type I and II error. It is necessary, if not pragmatic, to allow excursions in highly variable estuarine environments, but it does not necessarily indicate that the criteria are "overprotective" – just not always attainable. Not being overprotective is important because DO sags can be natural. The VP larval exposure model approach might be a way to translate exceedances into cumulative duration and have a stronger biological basis for "protectiveness" if there is uncertainty that even a moving average does not provide quantitative certainty of exposure stress and beneficial aquatic life use protection. This option notwithstanding, the combined use of moving averages on a 1-day, 7-day and 30-day basis and the targeted exceedance percentages is reasonable.

The second form of exceedance mentioned above is better addressed in this next section.

Whether Allowable Periods of Non-Compliance Are Scientifically Reasonable:

One focus of this work is on within-year allowable exceedance periods of noncompliance. Because natural background excursions from CCC and CMC are known to occur in productive marsh habitat, some exceedances are reasonable. Ten percent has traditionally been an EPA suggested rule of thumb when: 1) there is no other information upon which to base an allowable exceedance rate, 2) AND it was provided in consideration of grab samples. However, the 10% rule is not particularly recommended for use with dissolved oxygen assessments or used with high frequency assessment of dissolved oxygen conditions.

The reference based approach provides valuable insight as to the allowable exceedance frequency within years. The proposed reference-derived values (4% and 16% values for CMC and 7-day average CCC are more encouraging for use since they are based on best available site data. However, as pointed out in other states like in Delaware, some thought should be provided about the distribution of the criteria violations and what is your definition of "a violation event" (i.e., can you allow all the violations to occur on single long event? - see Tidal Murderkill River guidance text in Attachment 2). To count violations, must they be separated by a period of time to be considered as separate events? Then are you allowing that many events in a year or season or assessment period (i.e., like 3 yrs or 5 yrs or some other assessment). Another consideration is the degree or intensity of the exceedance, e.g., severe hypoxia or anoxia can be immediately lethal, but if the condition is of short duration, it would not constitute an exceedance under the percent criteria.

Here again, the stringency of criteria is dependent on monitoring program specifics and treatment of the data; blocking versus moving average will make a difference in how exceedance frequency is applied. This is where the understanding of the behavior of dissolved oxygen in the reference system versus impacted system is important. In the reference system, the diel cycle might occasionally dip below a threshold value in a random fashion. Diel hypoxic events differ from your impacted sloughs where there is an extended period of low dissolved oxygen. We return to the derivation of exceedance rate such that if you are using 30 day blocks – a single long event that might last for 2-3 weeks could register as one violation of the mean. By comparison, a daily time step of the rolling 30 day mean will register many violations. Therefore, your expectation on how the system behaves and your expected measure of protection should again be consistent in terms of how you view allowable exceedances (e.g. based on means computed from 15 minute interval data) and the subsequent monitoring and assessment approach (i.e., also from 15 minute interval data).

Finally, biological monitoring (fish and benthic macroinvertebrates) can be used as an additional line of evidence to assess status of impairment and the allowable noncompliance rate over a period of years. Community integrity has been used in freshwater and estuarine ecosystems as integrated measures reflecting habitat health (Karr 1981, Weisberg et al. 1997, Alden et al. 2002). The biology should inform the decision on how stable the desired communities are relative to the frequency of years of noncompliance. As referred to in the earlier section on recovery rates, the abilities of ecosystems to recover differ greatly, and depend on the pollutant, the magnitude and duration of the exceedance, and the physical and biological features of the ecosystem. Documented studies of recoveries are relatively few, but some systems recover from small stresses in six weeks whereas other systems take more than ten years to recover from severe stress (U.S. EPA 2016). EPA highlights many system level measures of health returned to pre-impairment conditions in about 2 years and therefore a 1 in 3 year allowable exceedances; alternatively, the Borja et al. 2010 review pointed towards 5-6 year recovery rates such that a 1 in 5 or 1 in 6 year allowable exceedance rate could be considered scientifically supported. Recovery rates were outside the scope of our discussion, however, additional information on the consideration of recovery rates to guide your standard setting and decision rules is provided in Attachment 3.

Exceedances in Back-End Sloughs:

The Panel supports the concept that in back-end sloughs the CMC exceedances may occur multiple times in a month without adversely affecting aquatic life. The CMC does not consider cumulative or repetitive exposure that might result in a lower tolerance, or the degree of the exceedance (e.g., severe hypoxia or anoxia) that might result in acutely lethal conditions. However, if the CCC is maintained as a moving average, conditions protective and supportive of the CMC with respect to cumulative stress may be reasonably assured.

Critical Period for Monitoring Compliance:

A strategic monitoring plan should be developed that considers several factors to characterize DO conditions that might exist in time and space in Suisun Marsh, and assesses changes that may occur over time. Yes, if there are known characteristic periods of the year that experience hypoxic events, it makes sense to monitor at those times more intensively and at the locations where they are known to occur. One should be aware that the foundation processes that drive these patterns can change with restoration or some other human intervention, or climate change. Monitoring can be flexible, and adaptive, with scoping or reconnaissance monitoring used to survey the potential for DO criteria exceedances over a range of habitats, locations and time periods. Monitoring needs to be most intensive in areas and at times when the DO conditions are close to the criteria. If surveys show that there are never problems, infrequent checks may be enough; if it's always hypoxic, extensive monitoring similarly may be unnecessary. The worst DO resources conditions under natural conditions are likely to occur when temperatures are warm and inflow low, mid-July through mid-October.

Because weather can have an important effect on hypoxia (or human causes such as the pond draining), it is important to "bracket" those periods to ensure if an early start or late end of the hypoxic period is captured, and so enough data are available to make moving averages work. Ideally, year-round monitoring should be used initially until the regularity of hypoxic events is known, and the sampling regime can be set with an appropriate temporal buffer preceding and following the hypoxic event period can be included in the sampling strategy. As noted above, missing part of the hypoxia cycle in a static or block averaging period could give a misleading result. Likewise, applying an exceedance frequency based on an annual monitoring cycle may be erroneous if applied to a much shorter monitoring period.

Monitoring Interval:

In general, the Panel is supportive of a 15-minute sampling frequency. Sampling density in time will affect your estimate of the mean and uncertainty about the mean. As sample size increases, error is reduced on the estimate of the variance and monitoring results move towards near perfect knowledge of the mean of that high temporal density data stream. Data storage is cheap. We could collect measurements at 1 sec intervals. A key question here is what is the phenomenon we are most concerned about and is 15 minutes or some other interval suitable to detect the phenomenon? Further, what was the basis for the data in the studies when they were evaluated for 24-48 hours? Did they develop the results with hourly measurements or continuous second by second measurements? And lastly, how representative is the station of the surrounding waters and for how long? In the case of a water release from an impoundment, the impact may be extensive in area while a dip in DO under natural summer conditions may reflect local scale processes at work (hence a desire to have 2-3 sites in operation if possible).

This question really speaks to the definition applied in the water quality standard. What is the regulatory definition of an event that is associated with the definition of "impairment" of the designated use? (For reference again, the Tidal Murderkill River guidance in Attachment 2 provides some consideration for defining an event and how many events in a year represent an impairment of the use when using continuous DO monitoring data for assessment.) A daily mean can be estimated from one measurement albeit with large uncertainty or from sub-daily scale measurements with increased accuracy with increasing sampling intensity. Hourly measurements can track the diel cycle of DO, sub-hourly measurements can provide strong support for understanding hypoxic event duration. A 15-minute interval is likely reasonable to estimate a mean and provide diagnostic data about event durations and causes of fish kills, for example. Given the physics of estuarine system, the detection of short duration events are probably a reflection of very local processes and consistent records of low DO events would be indicative of the system at the brink of a threshold of concern. By contrast, detecting longer duration hypoxic events at a station given with the tidal dynamics of these sloughs likely reflects a larger mass of water is experiencing the low DO phenomenon and therefore more of a concern.

Collateral Parameters for Data Interpretation:

We recommend collecting temperature, conductivity, and depth as key parameters to collect, along with DO. Data loggers that provide temperature, salinity and DO are readily available, not too expensive and very reliable.

Q3 Please comment on whether additional lines of evidence are necessary to assess DO attainment (e.g. biological confirmation) and overall beneficial use support.

The analysis presents a viable path forward that is protective of aquatic life beneficial uses. However, as noted above, historical data are somewhat limiting and may not cover the full range of conditions in time and space that occur in the study area. The best prospects are to continue to sample for compliance, and use those data to evaluate and refine the criteria, if warranted. And, research into sensitivity of resident organisms along with continued fish surveys, may reveal more sensitive species, additional areas of concern, and changes due to climate or other anthropogenic factors associated with pollutant loading, hydrologic modifications or habitat change as well as combined effects of other stressors. In Chesapeake Bay, benthic macroinvertebrate community assessments support State's decisions on listing a management segment as impaired or not. If the science develops such that benthic community data for Suisun Marsh could provide reliable diagnostic results to separate healthy from degraded habitat conditions, it has been used in other systems to supplement attainment decisions.

Q4 Given available data in Suisun Marsh and lessons learned, identify if there are other analyses/data and refinements that could strengthen the site-specific objectives for Suisun Marsh, and aid derivation of the criteria in similar habitats in SF Bay.

Setting the criteria is very dependent on the needs of the living resources or the intended use. To that end, the elements used here to derive criteria (Virginian Province approach, larval recruitment model, reference system approach, biological monitoring as supporting data) are an excellent example of the application of DO criteria development concepts. Application of these criteria to other places cannot proceed without redoing the steps, because different systems with have a different list of species and different aspects to this interpretation.

Future research may serve to refine this approach in other systems. Better documentation of life stage will help clarify the seasonal requirements for DO. Modeling has been used to target site specific understanding in setting criteria (e.g. Chesapeake Bay has a variance for blackwater systems DO targets that are lower than systems not so affected by natural DOC load, e.g., review the Delaware tidal Murderkill River document for their modeling assessments of the system.). Supporting the research to develop the larval recruitment curve(s) and encouraging laboratory mesocosm experiments on data lacking for native species, particularly for chronic exposures, will address important west coast-wide data gaps (Sutula et al. 2012).

References

Alden, R.W. III, D.M. Dauer, J.A. Ranasinghe, L.C. Scott, and R.J. Llansó. 2002. Statistical verification of the Chesapeake Bay benthic index of biotic integrity. Environmetrics, 13:473 498.

Borja, A., D. M. Dauer, M. Elliott, and C.A. Simenstad. 2010. Medium and long-term recovery of estuarine and coastal ecosystems: patterns, rates and restoration effectiveness. Estuaries and Coasts, 33:1249-1260.

Boynton, W.R., J.M. Testa, C.L.S. Hodgkins, J.L. Humphrey, and M.A.C. Ceballos. 2014. Maryland Chesapeake Bay Water Quality Monitoring Program. Ecosystem Processes Component. Level one Report No. 31. Interpretive Report. August 2014. Tech. Report Series No. TS-665-14 of the University of Maryland Center for Environmental Science. UMCES-CBL 2014-051. http://www.gonzo.cbl.umces.edu/documents/water_quality/Level1Report31.pdf

Jassby, A.D., B.E. Cole, and J.E. Cloern. 1997. The design of sampling transects for characterizing water quality in estuaries: Estuarine, Coastal and Shelf Science, 45: 285-302.

Karr, J.R. 1981.Assessment of biotic integrity using fish communities. Fisheries. 66:21-27.

Moyle, P.B., A.D. Manfree, and P.L. Fiedler. 2014. Suisun Marsh: Ecological History and Possible Futures. University of California Press.

Scott, J.T. and B.E. Haggard. 2015. Evaluating the assessment methodology for the chlorophyll-*a* and Secchi transparency criteria at Beaver Lake, Arkansas. Prepared for the Beaver Watershed Alliance. White paper. http://www.beaverwatershedalliance.org/pdf/assessment-methodology-final-report.pdf

Sutula, M., H. Bailey, and S. Poucher. 2012. Science Supporting Dissolved Oxygen Objectives in California Estuaries. Technical Report 684. Southern California Coastal Water Research Project. Costa Mesa, CA.

Tetra Tech. 2017. DO Criteria Recommendations for Suisun Marsh. Prepared for the San Francisco Regional Water Quality Board. March 2017.

U.S. Environmental Protection Agency. 2003. *Ambient Water Quality Criteria for Dissolved Oxygen, Water Clarity and Chlorophyll a for the Chesapeake Bay and Its Tidal Tributaries.* EPA 903-R-03-002. U.S. Environmental Protection Agency, Region 3, Chesapeake Bay Program Office, Annapolis, MD.

U.S. Environmental Protection Agency. 2016. Guidelines for Deriving Numerical National Water Quality Criteria for the Protection of Aquatic Organisms and Their Uses. Office of Research and Development, Environmental Research Laboratories, Duluth, MN. PB85-22749. https://www.epa.gov/sites/production/files/2016-02/documents/guidelines-water-quality-criteria.pdf

Weisberg, S.B., J.A. Ranasinghe, D.M. Dauer, L.C. Schaffner, R.J. Diaz, and J.B. Frithsen. 1997. An estuarine benthic index of biotic integrity (B-IBI) for the Chesapeake Bay. Estuaries, 20:149-158.

Attachments

Attachment 1: An example of binomial assessment of water quality meeting or exceeding the criterion.

Assume you have limited resources. You are only able to send a crew out to sample 12 days a year. There is some form or randomization that would be needed here – for example, randomly pick a start date in January, suppose it is Jan 15th. You sample Jan 15th then sample the 15th of every month. Alternatively you could randomly select a day each month. Either way, including the random element supports the statistical integrity underlying the tests. However, now you have in this case 12 samples. You know from setting your criteria that the allowable exceedance rate is 5%. You want to test your sampling results against this allowable rate. If I sample 12 times, what is the probability that 1 daily mean will fail the criterion? What if I get 2 daily means below the criterion? We can build a table of binomial probabilities to meet any sampling effort:

Binomial probability b(y; 12, 0.05) looks like this:

Y = 'successes' or in this case, actual measured exceedances/failures of the criterion in your monitoring program.
12 is the number of samples being collected for this example.
0.05 equals the 5% expected allowable violation rate based on an imagined allowable exceedance rate from a study of reference systems and a proposed criterion.

This example uses a reasonable set of conditions and measurement effort that might be appropriate for an assessment program with limited resources.

If you wanted to set your allowable violations as satisfying the criterion with a p-value of 0.1, or something close to it, we can use the table of probabilities below to see where the probability of a result becomes unlikely.

Reading the table, if I take 12 samples, and I am expect that the DO conditions will provide no more than 5% of measures below the criterion because that is what my reference told me we could have, then we look at where we get a probability less than 0.1. What I see is that we can allow 0, 1 or 2 daily means that were collected at random. 2 violations could happen 10% of the time a random sample was collected from a population of values where we select 12 samples and have an expected violation rate of 5%. If we get 3 samples that violate the daily mean, the likelihood of that is 1 in 100. The likelihood of 4 sample violating the criterion and still collecting it from water that only violates its criterion 5% of the time is at least 1 in 1000 (Note, with rounding here you only see 0.00 probability of the event. But if you carried out more decimal places you would compute and see the very small probability of the event beyond 1 in 100) – a very rare and unlikely result with 12 samples. The decision would be that the system is no meeting its criterion. Similarly, if 4 or more samples violated the criterion it would be even more rare to get that result if you are truly sampling from a population that looks like reference conditions. The decision for 3 or more samples not meeting the criterion would that your system is not meeting its goal of no more than 5% exceedances for the year and would be a failure that year.

Y	The computed binomial probability of getting Y 'successes' from 12 samples (Bernoulli formula used for computing the binomial probability)
0	0.54
1	0.34
2	0.10
3	0.01
4	0.00
5	0.00
6	0.00
7	0.00
8	0.00
9	0.00
10	0.00
11	0.00
12	0.00

In this example we picked a small but no unusual sample size for natural resource assessment. It probably has low power, we can calculate that as another step if we wanted to. For our purposes here I just provide the example to show this is one viable test and not unlike some states that collect around 10 samples in a year or season to compare against a criterion.

Attachment 2: A subsample of assessment approaches and decision frames from across the United States with regard to evaluating compliance to DO criteria.

Many states are exploring a variety of assessment approaches for DO criteria across a variety of habitats. States highlight the importance of taking into account the time of day for their sampling to address the minimum DO criteria. One example is Minnesota that applies a daily minimum of 5 mg O_2/L to its cool and warm water fisheries and splits the year into two seasons; May through September and October through April. Their assessment for dissolved oxygen requires no more than ten percent of the measurements taken in either period can violate the standard. Furthermore, measurements must be taken before 9:00 am to be representative of minimal conditions. Similarly, Oklahoma has a criterion of 5 mg O_2/L for warm water aquatic communities, but decreases that to 4 mg O_2/L during June 16 to October 15. Impairment is cited if more than 10% of the samples are below the criterion or if more than 2 samples are below 2 mg O_2/L. Under this form of wording for impairment assessment caution is recommended basing such assessments on percent of samples such that sufficient numbers of samples are collected for representative assessments.

Kansas (2011) was considering a variety of options to updating their 5 mg O_2/L minimum DO criterion. Options included: 1) lowering the DO criterion to a 4 mg/L instantaneous minimum. 2) assessing DO as a chronic impairment with binomial statistics (10% allowance of exceedance), explicitly stating allowances accounting for natural conditions, 4) explicitly excluding applying the criterion to the deepest portions of lakes (i.e. hypolimnetic waters).

For Massachusetts, in estuaries, their analysts compare DO data to the appropriate criterion (depending on a waterbody's classification) for surface water and depth measurements. (The national criteria daily minima (1.0 mg O_2/L less than the 7-day mean) were set to protect against acute (mortality) of sensitive species and they were also designed to prevent significant episodes of continuous or regularly recurring exposures to dissolved oxygen at or near the lethal threshold. DWM analysts use this daily minimum deviation (1.0 mg O_2/L) from the criterion for impairment decisions.) If all DO data meet (i.e., are above) the criterion, DO is considered sufficient to support the *Aquatic Life Use*. The analyst must evaluate the frequency and duration of excursions (whether or not they exceed 10% of the measurements) as well as the magnitude of any excursions (i.e., >1.0 mg O_2/L below the criterion). DO is identified as a cause of impairment if data indicate frequent, prolonged and/or severe excursion(s) from the appropriate criterion.

The temporal resolution and spatial density of measurements are variously considered across the country. In Oklahoma for example, for lakes, volume and space are taken into account and impairment is claimed if more than 50% of the lake water column has a dissolved oxygen concentration less than 2 mg O_2/L or if 10% of the surface samples are below the 5/4 mg O_2/L criteria.

Avoiding some of the challenges of grab sampling approaches to address temporal issues of diel cycling in DO behavior, states are advancing the uses of continuous monitoring data assessments. Washington State notes "Continuous sampling throughout the day can provide the lowest daily DO values; however, single "grab" samples are also used to determine compliance" (Department of Ecology, State of Washington 2009). Missouri evaluates stream reaches and recommends continuous monitoring data assessments at representative points in the stream (Missouri DNR 2010). Note, Missouri lists a sample period of days, a number of locations and a number of years involved in supporting a decision on impairment. The recommended sample size needed to estimate average daily mean and minimum DO concentrations in each of Missouri's ecological drainage units (EDU) are as follows:

- Continuous DO data collection efforts should target a deployment period of 68 days during the summer sampling period (July 1 – September 30);
- Data should be collected at 2 locations on each reference reach;
- All reference reaches should be monitored; and,
- Three years (summers) of data should be collected at each site.

Statistically, if they are randomly choosing the start date within the season from the period that would allow 68 days of monitoring, this would add a level of integrity to their assessment.

Rhode Island saltwater DO criteria are evaluated on cumulative exposures of low DO with established minimum standards. Therefore, Rhode Island is also moving to a reliance on continuously collected saltwater DO data or data that can be correlated to continuous data. Data are not interpolated but considered based on site specific assessment representing a region of the estuary (RI State Office of Water Staff, Pers. Comm.). Grab samples or similar DO data may still be considered if it can be correlated to continuous data or is representative of a longer time period.

Delaware has recently adopted site-specific DO criteria for the tidal Murderkill River (see http://www.dnrec.delaware.gov/swc/wa/Documents/WAS/Murderkill%20River%20Reports/Updated%20Drafts/Proposed%20Site-specific%20Dissolved%20Oxygen%20Criteria%20for%20Tidal%20Murderkill%20River.pdf.

The criteria and the assessment of the standard are:

The tidal portion of the Murderkill River has criteria for a daily averages and a one hour-average minimum criteria. Where continuous data are available, it will be assessed as rolling averages for the one hour minimum criteria and simple arithmetic averages for the daily average.

- For the one hour calculations, events less than 24 hours apart will be considered a single event. Two or more events more than 24 hours apart in one season will be considered not supporting of the use.
- Daily average criteria will be simple daily averages of the continuous data for each day in the period. Because of the hydrodynamics of the system, violations can occur over multiple day periods caused solely by tide and weather events.
- Violations less than 3 days apart will be considered a single event. Two or more violations in a single year, of the daily average will be considered as not supporting the use.

Attachment 3: Additional thoughts on considerations for selecting an assessment period for the impairment decision.

New literature on ecosystem recovery rates is available and highlights longer recoveries from stress than those used in U.S. EPA (2003). Using ecosystem recovery rates as a basis for defining an appropriate assessment period (e.g., U.S. EPA 2003), the new literature would suggest a longer assessment period than 3 years may be warranted and supported. U.S. EPA (2003) states "EPA guidance recommends use of a 1 in 3 year maximum allowable excursion recurrence frequency – number of times conditions in water are worse than those specified by the concentration and duration components of a freshwater life criterion for a toxic chemical". A key basis for this recommendation that defined a decision rule within a 3-year assessment period context was a 1989 literature survey of over 150 studies looking at recovery rates of freshwater ecosystems from various kinds of natural disturbances and anthropogenic stressors. The vast majority of macroinvertebrate and fish metric endpoints recovered in 2 years or less. However, a more recent review on recovery rates specific to estuarine and coastal ecosystems was published by Borja et al. (2010) in Estuaries and Coasts. Borja et al. summarized results from 51 studies used to evaluate recovery patterns as a function of various stressors. To be fair, many of the studies cited by Borja et al. (2010) pertained to systems that had experienced long-term degradation from a variety of stressors, as opposed to episodic impacts associated with a single parameter. However, similar to the 1989 EPA review, some studies showed near-term (months to a few years) recoveries of certain taxonomic groups. However, the lower boundaries for the majority of studies (see Table 2 in Borja et al.) is frequently 2-3 years while central tendencies are longer, and many recoveries take 6 years or more. Following the basis for EPA's support for decision rules based on a 3 year assessment, and now using the same logic with the support of this estuarine synthesis of new research on recovery rates of living resources by Borja et al. (2010), support here would suggest that longer assessment periods of 5-6 years instead of 3 years are supported and may be warranted.

What this leads you to consider is the difference between meeting your regulatory obligations versus having a tracking indicator. Your EPA-regulatory obligation may be to report impairment status in, for example, 5 year blocks. However, for your agency and as a means of tracking changes in the system, 5 year blocks are a long timeline for management to wait for a trend to show itself. Assuming you need at least 3 points to show a trend, that would be 15 years to get your first three data points. In the interim, you can use a rolling 5 year assessment with an annual time step as your annual tracking indicator of change. In this way you maximize the benefits of both approaches to follow change in your system. You get the sensitivity of the annual update to inform your managers of change over time in short time steps coupled with the regulatory performance of your 5-year block assessments to fulfill your EPA Clean Water Act 303d list reporting requirements. Over time, you should be able to leverage the two forms of tracking to evaluate status and highlight change and progress.